Relaxing the Hypotheses of Symmetry and Time-Reversibility in Genome Evolutionary Models

Jacques M. Bahi, Christophe Guyeux, Antoine Perasso

British Journal of Mathematics & Computer Science
5(4): 439-455, 2015, Article no.BJMCS.2015.031
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Relaxing the Hypotheses of Symmetry and Time-Reversibility in Genome Evolutionary Models

Jacques M. Bahi[1], **Christophe Guyeux**[1] **and Antoine Perasso**[2*]

[1] *UMR6147 Institut FEMTO-ST, Université de Franche-Comté, 25030 Besançon, France.*
[2] *UMR6249 Chrono-environnement, Université de Franche-Comté, 25030 Besançon, France.*

Article Information
DOI: 10.9734/BJMCS/2015/14280
Editor(s):
(1) Carlo Bianca, Department of Mathematical Sciences, Polytechnic University of Turin, Torino, Italy.
(2) Paul Bracken, Department of Mathematics, University of Texas-Pan American Edinburg, TX 78539, USA.
Reviewers:
(1) Anonymous, Anhui University, China.
(2) Anonymous, International Islamic University Malaysia, Malaysia.
Complete Peer review History:
http://www.sciencedomain.org/review-history.php?iid=728&id=6&aid=6913

Original Research Article

Received: 25 September 2014
Accepted: 24 October 2014
Published: 14 November 2014

Abstract

Various genome evolutionary models have been proposed these last decades to predict the evolution of a DNA sequence over time, essentially described using a mutation matrix. By essence, all of these models relate the evolution of DNA sequences to the computation of the successive powers of the mutation matrix. To make this computation possible, hypotheses are assumed for the matrix, such as symmetry and time-reversibility, which are not compatible with mutation rates that have been recently obtained experimentally on genes $ura3$ and $can1$ of the Yeast *Saccharomyces cerevisiae*. In this work, authors investigate systematically the possibility to relax either the symmetry or the time-reversibility hypothesis of the mutation matrix, by investigating all the possible matrices of size 2×2 and 3×3. As an application example, the experimental study on the Yeast *Saccharomyces cerevisiae* has been used in order to deduce a simple mutation matrix, and to compute the future evolution of the rate purine/pyrimidine for $ura3$ on the one hand, and of the particular behavior of cytosines and thymines compared to purines on the other hand.

**Corresponding author: E-mail: antoine.perasso@univ-fcomte.fr*

Keywords: Genome evolutionary models; stochastic processes; discrete dynamical system; matrix calculus; nucleotides mutations.

2012 Mathematics Subject Classification: 15A51; 15A16; 60G99; 92D10; 92D15; 92D20

1 Introduction

Due to mutations or recombination, some variations occur in the frequency of each codon, and these codons are thus not uniformly distributed into a given genome. Since the late '60s, various genome evolutionary models have been proposed to predict the evolution of a DNA sequence as generations pass. Mathematical models allow the prediction of such an evolution, in such a way that statistical values observed in current genomes can be at least partially recovered from hypotheses on past DNA sequences. Moreover, it can be attractive to study the genetic patterns (blocs of more than one nucleotide: dinucleotides, trinucleotides...) that appear and disappear depending on mutation parameters.

A first model for genomes evolution has been proposed in 1969 by Thomas Jukes and Charles Cantor [1]. This first model is very simple, as it supposes that each nucleotide has the probability m to mutate to any other nucleotide, as described in the following mutation matrix,

$$\begin{pmatrix} * & m & m & m \\ m & * & m & m \\ m & m & * & m \\ m & m & m & * \end{pmatrix}.$$

In that matrix, the nucleotides are ordered as (A, C, G, T), so that for instance the coefficient in row 3, column 2 represents the probability that the nucleotide G mutates into a C during the next time interval, *i.e.*, $P(G \to C)$. As diagonal elements can be deduced by the fact that the sum of each row must be equal to 1, they are omitted here.

This first attempt has been followed up by Motoo Kimura [2], who has reasonably considered that transitions ($A \longleftrightarrow G$ and $T \longleftrightarrow C$) should not have the same mutation rate than transversions ($A \longleftrightarrow T$, $A \longleftrightarrow C$, $T \longleftrightarrow G$, and $C \longleftrightarrow G$), this model being refined by Kimura in 1981, with three constant parameters to make a distinction between natural $A \longleftrightarrow T$, $C \longleftrightarrow G$ and unnatural transversions, leading to:

$$\begin{pmatrix} * & c & a & b \\ c & * & b & a \\ a & b & * & c \\ b & a & c & * \end{pmatrix}.$$

Joseph Felsenstein [3] has then supposed that the nucleotides frequency depends on the kind of nucleotide A,C,T,G. Such a supposition leads to a mutation matrix of the form:

$$\begin{pmatrix} * & \pi_C & \pi_G & \pi_T \\ \pi_A & * & \pi_G & \pi_T \\ \pi_A & \pi_C & * & \pi_T \\ \pi_A & \pi_C & \pi_G & * \end{pmatrix}$$

with π_A, π_C, π_G, and π_T denoting the frequency of occurance of each nucleotide, respectively. Masami Hasegawa, Hirohisa Kishino, and Taka-Aki Yano [4] have generalized the models of [2] and [3], introducing in 1985 the following mutation matrix:

$$\begin{pmatrix} * & \alpha\pi_C & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & * & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & * & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & \alpha\pi_G & * \end{pmatrix}.$$

Mutation	$ura3$	$can1$
$T \to C$	4	4
$T \to A$	14	9
$T \to G$	5	5
$C \to T$	16	20
$C \to A$	40	21
$C \to G$	11	9
$A \to T$	8	4
$A \to C$	6	5
$A \to G$	0	1
$G \to T$	28	20
$G \to C$	9	12
$G \to A$	26	40
Transitions	46	65
Transversions	121	85

Table 1: Summary of sequenced $ura3$ and $can1$ mutations [12]

These efforts have been continued by Tamura, who proposed in [5, 6] a simple method to estimate the number of nucleotide substitutions per site between two DNA sequences, by extending the model of Kimura (1980). The idea is to consider a two-parameter method, for the case where a GC bias exists. Let us denote by π_{GC} the frequency of this dinucleotide motif. Tamura supposes that $\pi_G = \pi_C = \dfrac{\pi_{GC}}{2}$ and $\pi_A = \pi_T = \dfrac{1-\pi_{GC}}{2}$, which leads to the following rate matrix:

$$\begin{pmatrix} * & \kappa(1-\pi_{GC})/2 & (1-\pi_{GC})/2 & (1-\pi_{GC})/2 \\ \kappa\pi_{GC}/2 & * & \pi_{GC}/2 & \pi_{GC}/2 \\ (1-\pi_{GC})/2 & (1-\pi_{GC})/2 & * & \kappa(1-\pi_{GC})/2 \\ \pi_{GC}/2 & \pi_{GC}/2 & \kappa\pi_{GC}/2 & * \end{pmatrix}.$$

All these models are special cases of the GTR model [7], in which the mutation matrix has the form (using obvious notations):

$$\begin{pmatrix} * & f_{AC}\pi_C & f_{AG}\pi_G & f_{AT}\pi_T \\ f_{AC}\pi_A & * & f_{CG}\pi_G & f_{CT}\pi_T \\ f_{AG}\pi_A & f_{CG}\pi_C & * & \pi_T \\ f_{AT}\pi_A & f_{CT}\pi_C & \pi_G & * \end{pmatrix}.$$

Non-reversible and non-symmetric models have, for their part, been considered in practical inferences since at least a decade for phylogenetic studies, see for instance [8, 9, 10]. As they are more regarded for their interest in practical inference investigations than on the theoretical side, they will not be developed in this article.

Due to mathematical complexity, matrices theoretically investigated to model evolution of DNA sequences are thus limited either by the hypotheses of symmetry and time-reversibility or by the desire to reduce the number of parameters under consideration. These hypotheses allow their authors to solve theoretically the DNA evolution problem, for instance by computing directly the successive powers of their mutation matrix. However, one can wonder whether such restrictions on the mutation rates are realistic. Focusing on this question, we used in [11] a recent research work of Lang and Murray [12], in which the per-base-pair mutation rates of the Yeast *Saccharomyces cerevisiae* have been experimentally measured (see Table 1), allowing us to calculate concrete mutation matrices for genes $ura3$ and $can1$. We deduced in [11] that none of the existing genomes evolution models can fit such mutation matrices, implying the fact that some hypotheses must be relaxed, even if this relaxation implies less ambitious models: current models do not match with what really occurs in concrete genomes, at least in the case of this yeast. Having these considerations in mind, the data obtained by Lang and Murray have been used in [11] in order to predict the evolution of

the rates or purines and pyrimidines in the particular case of $ura3$. Mathematical investigations and numerical simulations have been proposed, focusing on this particular gene and its associated matrix of size 2×2 (purines *vs.* pyrimidines), and of size 3×3 (cytosines and thymines compared to purines). Instead of focusing on two particular matrices, this extension of [11] investigates systematically all the possible mutation matrices of sizes 2×2 and 3×3. Thus, the study is finalized in this article, by investigating all the possible cases, and discussing about their mathematical and biological relevance.

The remainder of this research work is organized as follows. First of all the case of mutation matrices of size 2×2 is recalled in Section 2 and applied to the $ura3$ gene taking into account purines and pyrimidines mutations. A simulation is then performed to compare this non reversible model to the classical symmetric Cantor model. The next sections deal with all the possible 6-parameters models of size 3×3. In Section 3, a complete theoretical study is led encompassing all the particular situations, whereas in Section 4 an illustrative example focusing on the evolution of the purines, cytosines, and thymines triplet is given for $ura3$. We finally conclude this work in Section 5.

2 General Model of Size 2×2

In this section, a first general genome evolution model focusing on purines versus pyrimidines is proposed, to illustrate the method and as a pattern for further investigations. This model is applied to the case of the yeast *Saccharomyces cerevisiae*.

2.1 A convergence result

Let R and Y denote respectively the occurrence frequency of purines and pyrimidines in a sequence of nucleotides, and $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the associated mutation matrix, with $a = P(R \to R)$, $b = P(R \to Y)$, $c = P(Y \to R)$, and $d = P(Y \to Y)$ satisfying

$$\begin{cases} a + b = 1, \\ c + d = 1, \end{cases} \tag{2.1}$$

and thus $M = \begin{pmatrix} a & 1-a \\ c & 1-c \end{pmatrix}$.

The initial probability is denoted by $P_0 = (R_0 \;\; Y_0)$, where R_0 and Y_0 denote respectively the initial frequency of purines and pyrimidines. So the occurrence probability at generation n is $P_n = P_0 M^n$, where $P_n = (R(n) \;\; Y(n))$ is a probability vector such that $R(n)$ (resp. $Y(n)$) is the rate of purines (resp. pyrimidines) after n generations. The following theorem states the time asymptotic behavior of the probabilit P_n.

We recall the following result was proved in [11]:

thm 2.1. *Consider a DNA sequence under evolution, whose mutation matrix is* $M = \begin{pmatrix} a & 1-a \\ c & 1-c \end{pmatrix}$ *with* $a = P(R \to R)$ *and* $c = P(Y \to R)$*.*

- *If* $a = 1, c = 0$*, then the frequencies of purines and pyrimidines do not change as the generation pass.*
- *If* $a = 0, c = 1$*, then these frequencies oscillate at each generation between* $(R_0 \;\; Y_0)$ *(even generations) and* $(Y_0 \;\; R_0)$ *(odd generations).*
- *Else the value* $P_n = (R(n) \;\; Y(n))$ *of purines and pyrimidines frequencies at generation* n *is convergent to the following limit:*

$$\lim_{n\to\infty} P_n = \frac{1}{c+1-a} \begin{pmatrix} c & 1-a \end{pmatrix}.$$

rem 2.1. *Note that the case $a \neq 1-c$, resp. $a \neq c$, translates the non symmetry property, resp. the time reversibility property.*

2.2 Numerical Application

For numerical application, we will consider mutations rates in the *ura3* gene of the Yeast *Saccharomyces cerevisiae*, as obtained by Gregory I. Lang and Andrew W.Murray [12] and summed up in Table 1. They have measured phenotypic mutation rates, indicating that the per-base pair mutation rate at *ura3* is $m = 3.0552 \times 10^{-7}$/generation for the whole gene.

For the majority of Yeasts they studied, *ura3* is constituted by 804 bp: 133 cytosines, 211 thymines, 246 adenines, and 214 guanines. So $R_0 = \frac{246+214}{804} \approx 0.572$, and $Y_0 = \frac{133+211}{804} \approx 0.428$. Using these values in the historical model of Jukes and Cantor [1], we obtain the evolution depicted in Figure 1.

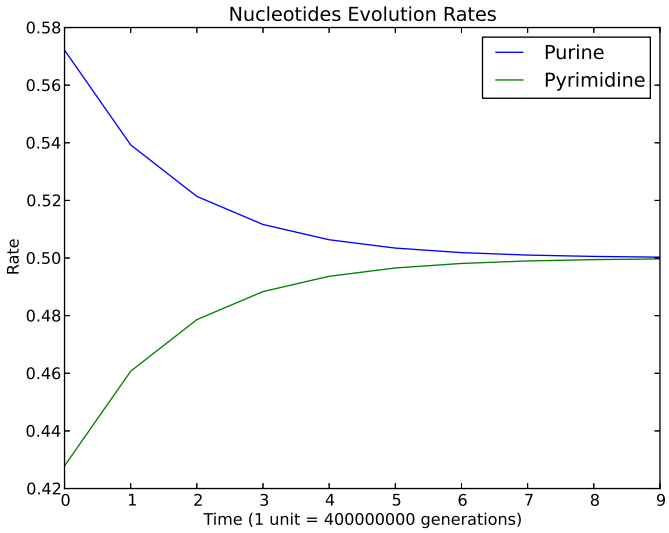

Figure 1: Prediction of purine/pyrimidine evolution of $ura3$ gene in symmetric Cantor model.

Theorem 2.1 allows us to compute the limit of the rates of purines and pyrimidines:

Computation of probability a**.** $P(R \to R) = (1-m) + P(A \to G)\frac{P_A(n)}{P_A(n)+P_G(n)} + P(G \to A)\frac{P_G(n)}{P_A(n)+P_G(n)}$. The use of Table 1 and the hypothesis that the base frequencies have already reached their steady states implies that $a = (1-m) + \left(m\frac{0}{46+121}\right) \times \frac{\frac{246}{804}}{\frac{246}{804}+\frac{214}{804}} + \left(m\frac{26}{46+121}\right) \times \frac{\frac{214}{804}}{\frac{246}{804}+\frac{214}{804}}$. We thus obtain that $a = 1 - \frac{17814m}{19205} \approx 0.999999716$.

Computation of probability c**.** Similarly, $P(Y \to Y) = (1-m) + P(C \to T)\frac{P_C}{P_C+P_T} + P(T \to C)\frac{P_T}{P_C+P_T} = (1-m) + m\frac{16}{46+121} \times \frac{133}{133+211} + m\frac{4}{46+121} \times \frac{211}{133+211} = 1 - m + m\frac{743}{14362}$.

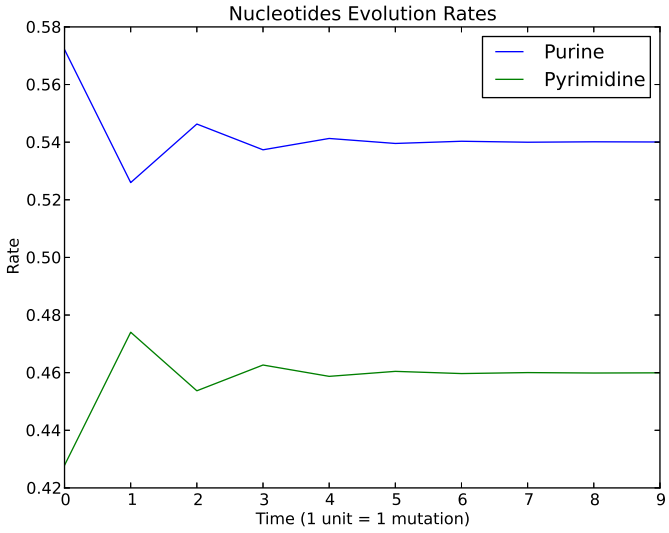

Figure 2: Prediction of purine/pyrimidine evolution of $ura3$ gene in general model of size 2×2.

So $c = P(Y \to R) = m\left(1 - \dfrac{743}{14362}\right) \approx 2.897 \times 10^{-7}$.

As a consequence the purine/pyrimidine mutation matrix that corresponds to the values of Table 1 is:

$$M = m \begin{pmatrix} \dfrac{1391}{19205} & \dfrac{17814}{19205} \\ \dfrac{13619}{14362} & \dfrac{743}{14362} \end{pmatrix}. \tag{2.2}$$

where $m = 3.0552 \times 10^{-7}$ as mentionned previously.

Using the value of m for the $ura3$ gene leads to $1 - a = 2.83391 \times 10^{-7}$ and $c = 2.89714 \times 10^{-7}$, which can be used in Theorem 2.1 to conclude that the rate of pyrimidines is convergent to 49.45% whereas the rate of purines converge to 50.55%. Numerical simulations using data published in [12] are given in Figure 2, leading to a similar conclusion.

3 A First Genomes Evolution Model of size 3×3 having 6 Parameters without Time-reversibility hypothesis

In order to investigate the evolution of the frequencies of cytosines and thymines in the gene $ura3$, a model of size 3×3 compatible with real mutation rates of the yeast *Saccharomyces cerevisiae* is now presented.

3.1 Formalization

Let us consider a line of yeasts where a given gene is sequenced at each generation, in order to clarify explanations. The $n-$th generation is obtained at time n, and the frequences of purines, cytosines,

and thymines at time n are respectively denoted by $P_R(n), P_C(n)$, and $P_T(n)$.

Let a be the probability that a purine is changed into a cytosine between two generations, that is: $a = P(R \to C)$. Similarly, denote by b, c, d, e, f the respective probabilities: $P(R \to T)$, $P(C \to R)$, $P(C \to T)$, $P(T \to R)$, and $P(T \to C)$. Contrary to existing approaches, $P(R \to C)$ is not supposed to be equal to $P(C \to R)$, and the same statement holds for the other probabilities. For the sake of simplicity, we will suppose in all that follows that a, b, c, d, e, f are not time dependent.

Let

$$M = \begin{pmatrix} 1-a-b & a & b \\ c & 1-c-d & d \\ e & f & 1-e-f \end{pmatrix}$$

be the mutation matrix associated to the probabilities mentioned above, and P_n the vector of occurrence, at time n, of each of the three kind of nucleotides. In other words, $P_n = (P_R(n)\ \ P_C(n)\ \ P_T(n))$. Under that hypothesis, P_n is a probability vector: $\forall n \in \mathbb{N}$,

- $P_R(n), P_C(n), P_T(n) \in [0, 1]$,
- $P_R(n) + P_C(n) + P_T(n) = 1$,

Let $P_0 = (P_R(0)\ \ P_C(0)\ \ P_T(0)) \in [0, 1]^3$ be the initial probability vector. We have obviously:

$$P_R(n+1) = P_R(n)P(R \to R) + P_C(n)P(C \to R) + P_T(n)P(T \to R),$$

with similar equalities for $P_C(n+1)$ and $P_T(n+1)$ so that

$$P_n = P_{n-1}M = P_0M^n. \tag{3.1}$$

In all that follows we wonder if, given the parameters a, b, c, d, e, f as in [12], one can determine the frequency of occurrence of any of the three kind of nucleotides when n is sufficiently large, in other words if the limit of P_n is accessible by computations.

3.2 Resolution

This section, that is a preliminary of the convergence study, is devoted to the determination of the powers of matrix M in the general case and some particular situations

3.2.1 Determination of M^n in the general case

The characteristic polynomial of M is equal to

$$\begin{aligned} \chi_M(x) &= x^3 + (s-3)x^2 + (p-2s+3)x - 1 + s - p \\ &= (x-1)\left(x^2 + (s-2)x + (1-s+p)\right), \end{aligned}$$

where

$$\begin{gathered} s = a + b + c + d + e + f, \\ p = ad + ae + af + bc + bd + bf + ce + cf + de, \\ det(M) = 1 - s + p. \end{gathered}$$

The discriminant of the polynomial of degree 2 in the factorization of χ_M is equal to $\Delta = (s-2)^2 - 4(1-s-p) = s^2 - 4p$. Let x_1 and x_2 the two roots (potentially complex or equal) of χ_M, given by

$$x_1 = \frac{-s+2-\sqrt{s^2-4p}}{2} \text{ and } x_2 = \frac{-s+2+\sqrt{s^2-4p}}{2}. \tag{3.2}$$

Let $n \in \mathbb{N}, n \geqslant 2$. As χ_M is a polynomial of degree 3, a division algorithm of X^n by $\chi_M(X)$ leads to the existence and uniqueness of two polynomials Q_n and R_n, such that

$$X^n = Q_n(X)\chi_2(X) + R_n(X), \tag{3.3}$$

where the degree of R_n is lower than or equal to the degree of χ_M, *i.e.*, $R_n(X) = a_n X^2 + b_n X + c_n$ with $a_n, b_n, c_n \in \mathbb{R}$ for every $n \in \mathbb{N}$. By evaluating (3.3) in the three roots of χ_M, we find the system

$$\left\{ \begin{array}{ll} 1 & = a_n + b_n + c_n \\ x_1^n & = a_n x_1^2 + b_n x_1 + c_n \\ x_2^n & = a_n x_2^2 + b_n x_2 + c_n \end{array} \right.$$

This system is equivalent to

$$\left\{ \begin{array}{lllll} c_n & + & b_n & + & a_n & = 1 \\ & & b_n(x_1 - 1) & + & a_n(x_1^2 - 1) & = x_1^n - 1 \\ & & b_n(x_2 - 1) & + & a_n(x_2^2 - 1) & = x_2^n - 1 \end{array} \right.$$

If we suppose that $x_1 \neq 1$, $x_2 \neq 1$, and $x_1 \neq x_2$, then standard algebraic computations give

$$\left\{ \begin{array}{l} a_n = \dfrac{1}{x_2 - x_1}\left[\dfrac{x_2^n - 1}{x_2 - 1} - \dfrac{x_1^n - 1}{x_1 - 1}\right], \\ \\ b_n = \dfrac{x_1 + 1}{x_1 - x_2}\dfrac{x_2^n - 1}{x_2 - 1} + \dfrac{x_2 + 1}{x_2 - x_1}\dfrac{x_1^n - 1}{x_1 - 1}, \\ \\ c_n = 1 - a_n - b_n. \end{array} \right.$$

Using for $i = 1, 2$ and $n \in \mathbb{N}$ the following notation,

$$X_i(n) = \frac{x_i^n - 1}{x_i - 1}, \tag{3.4}$$

and since $x_2 - x_1 = \sqrt{\Delta}$, the system above can be rewritten as

$$\left\{ \begin{array}{l} a_n = \dfrac{X_2(n) - X_1(n)}{\sqrt{\Delta}}, \\ \\ b_n = \dfrac{(x_2 + 1)X_1(n) - (x_1 + 1)X_2(n)}{\sqrt{\Delta}}, \\ \\ c_n = 1 + \dfrac{x_1 X_2(n) - x_2 X_1(n)}{\sqrt{\Delta}}. \end{array} \right. \tag{3.5}$$

By evaluating (3.3) in M and due to the theorem of Cayley-Hamilton, we finally have for every integer $n \geqslant 1$,

$$M^n = a_n M^2 + b_n M + c_n I_3, \tag{3.6}$$

where I_3 is the identity matrix of size 3, a_n, b_n, and c_n are given by (3.5), and M^2 is given by

$$M^2 = \left(\begin{array}{c|c|c} \begin{array}{c} a^2 + 2ab + ac - 2a \\ +b^2 + be - 2b + 1 \end{array} & \begin{array}{c} -a^2 - ab - ac \\ -ad + 2a + bf \end{array} & \begin{array}{c} -ab + ad - b^2 \\ -be - bf + 2b \end{array} \\ \hline \begin{array}{c} -ac - bc - c^2 \\ -cd + 2c + de \end{array} & \begin{array}{c} ac + c^2 + 2cd - 2c \\ +d^2 + df - 2d + 1 \end{array} & \begin{array}{c} bc - cd - d^2 \\ -de - df + 2d \end{array} \\ \hline \begin{array}{c} -ae - be + cf \\ -e^2 - ef + 2e \end{array} & \begin{array}{c} ae - cf - df \\ -ef - f^2 + 2f \end{array} & \begin{array}{c} be + df + e^2 + 2ef \\ -2e + f^2 - 2f + 1 \end{array} \end{array} \right).$$

3.2.2 Determination of M^n in particular situations

Formulations of (3.5) only hold for $x_1 \neq x_2$, $x_1 \neq 1$, and $x_2 \neq 1$. We now investigate these latter cases.

Preliminaries Let us firstly remark that, as the mutation matrix M is stochastic, we have necessarily $0 \leqslant a+b \leqslant 1$, $0 \leqslant c+d \leqslant 1$, and $0 \leqslant e+f \leqslant 1$. These inequalities imply that $s \in [0,3]$. Consequently from the definition of p one can check that $p = ad + a(e+f) + b(c+d) + bf + c(e+f) + de \leqslant ad + a + b + bf + c + de \leqslant s$, as each parameter is in $[0,1]$. To sum up,

$$0 \leqslant p \leqslant s \leqslant 3. \tag{3.7}$$

Suppose now that $\Delta \geqslant 0$. Then (3.2) and (3.7) imply that

$$x_1 = \frac{-s+2-\sqrt{\Delta}}{2} \in [-2;1], x_2 = \frac{-s+2+\sqrt{\Delta}}{2} \in \left[-\frac{1}{2};\frac{5}{2}\right] \tag{3.8}$$

Note that, as we deal with a stochastic process, the module of the eigenvalues of M are smaller than 1, so $|x_1| \leqslant 1$ and $|x_2| \leqslant 1$.

Suppose that $x_1 = 1$ Then $-s = \sqrt{s^2-4p} \iff s = p = 0$. So $a = b = c = d = e = f = 0$, and the mutation matrix is equal to the identity of size 3. Conversely, if $a = b = c = d = e = f = 0$, then $x_1 = 1$.

In that situation, the system does not evolve.

Suppose that $x_2 = 1$ (and $x_1 \neq 1$) Then $s = \sqrt{s^2-4p} \iff p = 0$. In that situation, $x_1 = 1-s$ and 1 is root of multiplicity 2 of χ_2, whereas $x_1 = 1-s$ is its third root. As the case $x_1 = 1$ has already been regarded, we can consider that $s \neq 0$. Using (3.3), These facts lead to the following system:

$$\begin{cases} 1 &= a_n + b_n + c_n, \\ n &= 2a_n + b_n, \\ (1-s)^n &= (1-s)^2 a_n + (1-s)b_n + c_n. \end{cases}$$

Standard computations then give the following formula:

$$\begin{cases} a_n = \dfrac{-1+sn+(1-s)^n}{s^2}, \\ b_n = \dfrac{(3-s)+(s^2-2s)n+(s-3)(1-s)^n}{s}, \\ c_n = \dfrac{(s-1)(2s-1)-s(s-1)^2 n-(s^2-3s+1)(1-s)^n}{s^2}. \end{cases} \tag{3.9}$$

Case $x_1 = x_2 \neq 1$ ($\Delta = 0$) Then (3.8) implies that $x_1 = 1-s/2 \in \left[-\frac{1}{2},1\right)$. From a differentiation of (3.3) one deduces that x_1 satisfies the following system for every $n \in \mathbb{N}^*$,

$$\begin{cases} 1 &= a_n + b_n + c_n \\ x_1^n &= a_n x_1^2 + b_n x_1 + c_n \\ n x_1^{n-1} &= 2a_n x_1 + b_n \end{cases}$$

Standard algebraic computations give, since $x_1 \neq 1$,

$$\begin{cases} a_n = n\dfrac{x_1^{n-1}}{x_1 - 1} - \dfrac{X_1(n)}{x_1 - 1} \\ \\ b_n = X_1(n) - a_n(x_1 + 1) \\ \\ c_n = 1 - a_n - b_n \end{cases} \tag{3.10}$$

where $X_1(n)$ is defined in (3.4).

3.3 Convergence study

3.3.1 Convergence study in the general case

We suppose in this section that $x_1 \neq x_2$, $x_1 \neq 1$, and $x_2 \neq 1$. So formulations of (3.5) hold for a_n, b_n, and c_n. We split the study convergence in several sub-cases, that are the objects of Theorems 3.1-3.5.

thm 3.1. *Suppose that $|x_1| < 1$ and $|x_2| < 1$. Then the frequencies $P_R(n), P_C(n)$, and $P_T(n)$ of occurrence at time n of purines, cytosines, and thymines in the considered gene, converge to the following values:*

- $P_R(n) \longrightarrow \dfrac{ce + cf + de}{p - bf + df}$
- $P_C(n) \longrightarrow \dfrac{ae + af + bf}{p - bf + df}$
- $P_T(n) \longrightarrow \dfrac{ad + bc + bd}{p - bf + df}$

Proof. If $|x_1| < 1$ and $|x_2| < 1$ then $X_i(n) \longrightarrow \dfrac{1}{1 - x_i}$ for $i = 1, 2$ and so

$$a_n \longrightarrow \frac{1}{\sqrt{\Delta}} \left(\frac{1}{1 - x_2} - \frac{1}{1 - x_1} \right).$$

Denote by a_∞ this limit. We have

$$a_\infty = \frac{x_2 - x_1}{\sqrt{\Delta}(1 - x_2)(1 - x_1)} = \frac{1}{(1 - x_2)(1 - x_1)} = \frac{1}{\dfrac{s + \sqrt{\Delta}}{2}\dfrac{s - \sqrt{\Delta}}{2}},$$

and finally

$$a_\infty = \frac{4}{s^2 - \Delta} = \frac{1}{p}.$$

Similarly, $b_n = X_1(n) - a_n(x_1 + 1)$ satisfies

$$b_n \longrightarrow \frac{1}{1 - x_1} - \frac{x_1 + 1}{p}.$$

The following computations

$$\frac{1}{1 - x_1} = \frac{2}{s + \sqrt{\Delta}} = \frac{2(s - \sqrt{\Delta})}{s^2 - \Delta} = \frac{s - \sqrt{\Delta}}{2p},$$
$$\frac{x_1 + 1}{p} = \frac{-s + 4 - \sqrt{\Delta}}{2p},$$

finally yield

$$b_\infty = \frac{s-2}{p}.$$

So

$$c_n \longrightarrow 1 - a_\infty - b_\infty = \frac{p-s+1}{p},$$

and to sum up, the distribution limit is given by

$$\begin{cases} a_\infty = \dfrac{1}{p} \\ b_\infty = \dfrac{s-2}{p} \\ c_\infty = \dfrac{p-s+1}{p} \end{cases} \tag{3.11}$$

Using the latter values in (3.6), we can determine the limit of M^n, which is $a_\infty M^2 + b_\infty M + c_\infty I_3$. All computations done, we find the following limit for M^n,

$$\frac{1}{p-bf+df}\begin{pmatrix} ce+cf+de & ae+af+bf & ad+bc+bd \\ ce+cf+de & ae+af+bf & ad+bc+bd \\ ce+cf+de & ae+af+bf & ad+bc+bd \end{pmatrix}.$$

Using (3.1), we can thus finally determine the limit of $P_n = P_0 M^n = (P_R(0)\ \ P_C(0)\ \ P_T(0))M^n$. □

thm 3.2. *Suppose that $|x_1| = 1, x_1 \neq 1$, and $|x_2| \neq 1$. Then the evolutionary model is not convergent. More precisely, we have:*

- $P_R(2n) = (a^2+2ab+ac-2a+b^2+be-2b+1)P_R(0) + (-a^2-ab-ac-ad+2a+bf)P_C(0) + (-ab+ad-b^2-be-bf+2b)P_T(0)$,
- $P_R(2n+1) = (1-a-b)P_R(0) + aP_C(0) + bP_T(0)$,
- $P_C(2n) = (-ac-bc-c^2-cd+2c+de)P_R(0) + (ac+c^2+2cd-2c+d^2+df-2d+1)P_C(0) + (bc-cd-d^2-de-df+2d)P_T(0)$,
- $P_C(2n+1) = cP_R(0) + (1-c-d)P_C(0) + dP_T(0)$,
- $P_T(2n) = (-ae-be+cf-e^2-ef+2e)P_R(0) + (ae-cf-df-ef-f^2+2f)P_C(0) + (be+df+e^2+2ef-2e+f^2-2f+1)P_T(0)$,
- $P_T(2n+1) = eP_R(0) + fP_C(0) + (1-e-f)P_T(0)$,

Proof. Suppose that $|x_1| = 1$ and $|x_2| \neq 1$. Then $x_1, x_2 \in \mathbb{R}$, and so $x_1 = 1$ or $x_1 = -1$. The first case has yet been regarded.

If $x_1 = -1$, then $-s+2-\sqrt{\Delta} = -2$ (due to (3.2)). So $s = 4-\sqrt{\Delta}$, and so $s^2-4p = 4-4s+s^2$. Consequently, $p = s-1$. But $x_1x_2 = 1-s+p$, so $x_1x_2 = 0$, which leads to $x_2 = 0$. Using (3.5), we can thus conclude that $a_n = 1 - \dfrac{(-1)^n-1}{-2} = \dfrac{1+(-1)^n}{2}$. So $a_{2n} = 1$ and $a_{2n+1} = 0$. Similarly, $b_{2n} = 0$ and $b_{2n+1} = 1$, and finally $c_n = 0, \forall n \in \mathbb{N}$.

These values for a_n, b_n, and c_n lead to the following values for M^n:

$$\begin{cases} M^{2n} = M^2 \\ M^{2n+1} = M. \end{cases}$$

□

rem 3.1. *The case* $|x_1| \neq 1$ *and* $|x_2| = 1$ *necessarilly implies that* $x_2 = 1$*, which is in contradiction with the assumptions made in preamble of Section 3.3.1.*

thm 3.3. *If* $|x_1| = |x_2|$*, but* $x_1, x_2 \in \mathbb{C}\backslash\mathbb{R}$*, then* $(P_R(n)\ P_C(n)\ P_T(n)) = (P_R(0)\ P_C(0)\ P_T(0)) \times (a_n M^2 + b_n M + c_n I_3)$*, where*

- $a_n = -\dfrac{sin\left(\frac{n\theta}{2}\right) sin\left(\frac{(n-1)\theta}{2}\right)}{sin\left(\frac{\theta}{2}\right) sin(\theta)}$,
- $b_n = \dfrac{2\ sin\left(\frac{n\theta}{2}\right) sin\left(\frac{(n-2)\theta}{2}\right)\ cos\left(\frac{\theta}{2}\right)}{sin(\theta)\ sin\left(\frac{\theta}{2}\right)}$,
- $c_n = 1 - \dfrac{sin\left(\frac{n\theta}{2}\right) sin\left(\frac{(n-3)\theta}{2}\right)}{sin(\theta)\ sin\left(\frac{\theta}{2}\right)}$.

with $e^{-i\theta} = x_1$.

Proof. Suppose that $|x_1| = |x_2|$, but $x_1, x_2 \in \mathbb{C}\backslash\mathbb{R}$. Then x_1 and x_2 are complex and conjugate, of the form $x_1 = e^{-i\theta}$, $x_2 = e^{i\theta}$, with $\theta \not\equiv 0[\pi]$. So $x_1 - x_2 = \sqrt{\Delta} = e^{-i\theta} - e^{i\theta} = -2i\ sin(\theta)$, and

$$
\begin{aligned}
a_n &= \frac{X_2(n) - X_1(n)}{\sqrt{\Delta}} = \frac{X_1(n) - X_2(n)}{2i\ sin(\theta)} \\
2i\ sin(\theta)\ a_n &= \frac{e^{-in\theta} - 1}{e^{-i\theta} - 1} - \frac{e^{in\theta} - 1}{e^{i\theta} - 1} \\
&= \frac{e^{-in\frac{\theta}{2}}}{e^{-i\frac{\theta}{2}}} \frac{e^{-in\frac{\theta}{2}} - e^{in\frac{\theta}{2}}}{e^{-i\frac{\theta}{2}} - e^{i\frac{\theta}{2}}} - \frac{e^{in\frac{\theta}{2}}}{e^{i\frac{\theta}{2}}} \frac{e^{in\frac{\theta}{2}} - e^{-in\frac{\theta}{2}}}{e^{i\frac{\theta}{2}} - e^{-i\frac{\theta}{2}}} \\
&= e^{-i\frac{(n-1)\theta}{2}} \frac{-2i\ sin\left(\frac{n\theta}{2}\right)}{-2i\ sin\left(\frac{\theta}{2}\right)} - e^{i\frac{(n-1)\theta}{2}} \frac{2i\ sin\left(\frac{n\theta}{2}\right)}{2i\ sin\left(\frac{\theta}{2}\right)} \\
&= \frac{sin\left(\frac{n\theta}{2}\right)}{sin\left(\frac{\theta}{2}\right)} \left(e^{-i\frac{(n-1)\theta}{2}} - e^{i\frac{(n-1)\theta}{2}}\right).
\end{aligned}
$$

Finally,

$$
a_n = -\frac{sin\left(\frac{n\theta}{2}\right) sin\left(\frac{(n-1)\theta}{2}\right)}{sin\left(\frac{\theta}{2}\right) sin(\theta)}.
$$

Similarly,

$$
\begin{aligned}
\sqrt{\Delta} b_n &= (x_2 + 1)X_1(n) - (x_1 + 1)X_2(n) \\
-2i\ sin(\theta) b_n &= \left(e^{i\theta} + 1\right) e^{-i\frac{(n-1)\theta}{2}} \frac{sin\left(\frac{n\theta}{2}\right)}{sin\left(\frac{\theta}{2}\right)} - \left(e^{-i\theta} + 1\right) e^{i\frac{(n-1)\theta}{2}} \frac{sin\left(\frac{n\theta}{2}\right)}{sin\left(\frac{\theta}{2}\right)} \\
&= \frac{sin\left(\frac{n\theta}{2}\right)}{sin\left(\frac{\theta}{2}\right)} \left[e^{-i\frac{(n-3)\theta}{2}} + e^{-i\frac{(n-1)\theta}{2}} - e^{i\frac{(n-3)\theta}{2}} - e^{i\frac{(n-1)\theta}{2}}\right] \\
b_n &= \frac{sin\left(\frac{n\theta}{2}\right)}{sin(\theta) sin\left(\frac{\theta}{2}\right)} \left(sin\left(\frac{(n-3)\theta}{2}\right) + sin\left(\frac{(n-1)\theta}{2}\right)\right).
\end{aligned}
$$

and finally,

$$b_n = \frac{2\ sin\left(\frac{n\theta}{2}\right) sin\left(\frac{(n-2)\theta}{2}\right)\ cos\left(\frac{\theta}{2}\right)}{sin(\theta)\ sin\left(\frac{\theta}{2}\right)}.$$

As $c_n = 1 - a_n - b_n$, we have:

$$c_n = 1 - \frac{sin\left(\frac{n\theta}{2}\right) sin\left(\frac{(n-3)\theta}{2}\right)}{sin(\theta)\ sin\left(\frac{\theta}{2}\right)}.$$

□

3.3.2 Convergence study in particular situations

The case where $x_1 = 1$ has already been discussed, it implies that $a = b = c = d = e = f = 0$, and so the system does not evolve. The other particular situations are invastigated in the two following theorems.

thm 3.4. *Suppose that $x_2 = 1$ and $x_1 \neq 1$ (or equivalently $p = 0$). Then the system is well formulated if and only if $M^2 + s(s-2)M - (s-1)^2 I_3 \neq 0$. In that situation, we have:*

- *either $s \in]0, 2[$, and so $(P_R(n)\ \ P_C(n)\ \ P_T(n)) \longrightarrow (P_R(0)\ \ P_C(0)\ \ P_T(0)) \times \frac{1}{s^2}[-M^2 + s(3-s)M + (s-1)(2s-1)I_3]$.*
- *or $s = 2$, and so $(P_R(2n)\ \ P_C(2n)\ \ P_T(2n)) \longrightarrow (P_R(0)\ \ P_C(0)\ \ P_T(0))$ whereas $(P_R(2n+1)\ \ P_C(2n+1)\ \ P_T(2n+1)) (P_R(0)\ \ P_C(0)\ \ P_T(0)) \times (-2M^2 + 4M + 2I_3)$.*

Proof. Using (3.9), we can deduce that M^n is equal to:

$$\begin{aligned} a_n M^2 + b_n M + c_n I_3 \quad = \quad & \frac{n}{s}[M^2 + s(s-2)M - (s-1)^2 I_3] \\ & + \frac{1}{s^2}(1-s)^n[M^2 + s(s-3)M - (s^2 - 3s + 1)I_3] \\ & + \frac{1}{s^2}[-M^2 + s(3-s)M + (s-1)(2s-1)I_3]. \end{aligned}$$

Several cases can be deduced from this equality.

- If $s \in]0, 2[$, then M^n is bounded if and only if $M^2 + s(s-2)M - (s-1)^2 I_3 = 0$. In that condition, $M^n \longrightarrow \frac{1}{s^2}[-M^2 + s(3-s)M + (s-1)(2s-1)I_3]$.
- If $s = 2$, then another time M^n is bounded if and only if $M^2 + s(s-2)M - (s-1)^2 I_3 = 0$. In that condition, $M^{2n} \longrightarrow I_3$, whereas $M^{2n+1} \longrightarrow -2M^2 + 4M + 2I_3$.
- Finally, if $s > 2$, then as $s = a + b + c + d + e + f$ and $a, b, c, d, e, f \in [0, 1]$, we have necessarily at least three coefficients in a, b, c, d, e, f that are non zero. So at least one product in $abc, abd, abe, abf, acd, ace, acf, ade, adf, aef, bcd, bce, bcf, bde, bdf, bef, cde, cdf, cef, def$ is strictly positive. This is impossible, as $p = ad + ae + af + bc + bd + bf + ce + cf + de$ is equal to 0.

□

thm 3.5. *Suppose that $x_1 = x_2 \neq 1$ (or equivalently $s^2 = 4p$). Then the probabilities $P_R(n), P_C(n)$, and $P_T(n)$ of occurrence at time n of a purine, cytosine, and thymine on the considered nucleotide, converge to the following values:*

- $P_R(n) \longrightarrow \frac{4}{s^2}(ce + cf + de)$,

- $P_C(n) \longrightarrow \frac{4}{s^2}(ae + af + bf)$,
- $P_T(n) \longrightarrow \frac{4}{s^2}(ad + bc + bd)$.

Proof. In that case $\Delta = 0$, meaning that (3.10) holds. Since $x_1 \in \left[-\frac{1}{2}, 1\right)$, one gets the following limits,

$$\lim_{n\to\infty} X_1(n) = -\frac{1}{1-x_1},$$
$$\lim_{n\to\infty} x_1^n = 0, \lim_{n\to\infty} n x_1^{n-1} = 0,$$

and finally (a_n, b_n, c_n) converges to $(a_\infty, b_\infty, c_\infty)$ with

$$\begin{cases} a_\infty = \dfrac{1}{(1-x_1)^2} = \dfrac{4}{s^2} \\ b_\infty = \dfrac{-2x_1}{(1-x_1)^2} = 4\,\dfrac{s-2}{s^2} \\ c_\infty = \dfrac{2x_1-1}{(1-x_1)^2} + 1 = \left(1 - \dfrac{2}{s}\right)^2 \end{cases}$$

Using these values in (3.6), we can determine the limit of M^n, which is $a_\infty M^2 + b_\infty M + c_\infty I_3$, where I_3 is the identity matrix of size 3. All computations done, we find

$$M^n \longrightarrow \frac{4}{s^2}\begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{pmatrix}$$

with $M_{11} = \dfrac{s^2}{4} - p + ce + cf + de$, $M_{12} = ae + af + bf$, $M_{13} = ad + bc + bd$, $M_{21} = ce + cf + de$, $M_{22} = \dfrac{s^2}{4} - p + ae + af + bf$, $M_{23} = ad + bd + bc$, $M_{31} = ce + de + cf$, $M_{32} = ae + af + bf$, and $M_{33} = \dfrac{s^2}{4} - p + ad + bc + bd$. However, since $x_1 = x_2$, we have $\Delta = s^2 - 4p = 0$ and so

$$M^n \longrightarrow \frac{4}{s^2}\begin{pmatrix} ce + cf + de & ae + af + bf & ad + bc + bd \\ ce + cf + de & ae + af + bf & ad + bc + bd \\ ce + cf + de & ae + af + bf & ad + bc + bd \end{pmatrix},$$

□

4 Application in Concrete Genomes Prediction

We consider another time the numerical values for mutations published in [12]. Gene *ura3* of the Yeast *Saccharomyces cerevisiae* has a mutation rate of 3.80×10^{-10}/bp/generation [12]. As this gene is constituted by 804 nucleotides, we can deduce that its global mutation rate per generation is equal to $m = 3.80 \times 10^{-10} \times 804 = 3.0552 \times 10^{-7}$. Let us compute the values of a, b, c, d, e, and f. The first line of the mutation matrix is constituted by $1 - a - b = P(R \to R)$, $a = P(R \to T)$, and $b = P(R \to C)$. $P(R \to R)$ takes into account the fact that a purine can either be preserved (no mutation, probability $1 - m$), or mutate into another purine ($A \to G$, $G \to A$). As the generations pass, authors of [12] have counted 0 mutations of kind $A \to G$, and 26 mutations of kind $G \to A$. Similarly, there were 28 mutations $G \to T$ and 8: $A \to T$, so 36: $R \to T$. Finally, 6: $A \to C$ and 9: $G \to C$ lead to 15: $R \to C$ mutations. The total of mutations to consider when evaluating the first line

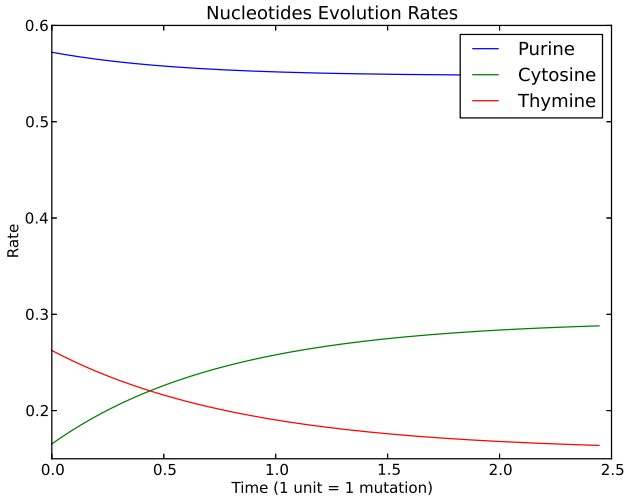

Figure 3: Prediction of evolution concerning the purine, thymine, and cytosine rates in *ura3*. Non-symmetric Model of size 3×3.

is so equal to 77. All these considerations lead to the fact that $1-a-b=(1-m)+m\frac{26}{77}$, $a=\frac{36m}{77}$, and $b=\frac{15m}{77}$. A similar reasoning leads to $c=\frac{19m}{23}$, $d=\frac{4m}{23}$, $e=\frac{51m}{67}$, and $f=\frac{16m}{67}$.

In that situation, $s=a+b+c+d+e+f=\frac{205m}{77}\approx 8.134\times 10^{-7}$, and $p=\frac{207488m^2}{118657}\approx 1.632\times 10^{-13}$. So $\Delta=s^2-4p=\frac{854221m^2}{9136589}>0$, $x_1=1-\frac{m}{2}\left(\frac{205}{77}+\sqrt{\frac{854221}{9136589}}\right)$, and $x_2=1-\frac{m}{2}\left(\frac{205}{77}-\sqrt{\frac{854221}{9136589}}\right)$. As $x_1\approx 0.9999685\in[0,1]$ and $x_2\approx 0.9999686\in[0,1]$, we have, due to Theorem 3.1:

- $P_R(n)\longrightarrow\frac{ce+cf+de}{p-bf+df}$
- $P_C(n)\longrightarrow\frac{ae+af+bf}{p-bf+df}$
- $P_T(n)\longrightarrow\frac{ad+bc+bd}{p-bf+df}$

Using the data of [12], we find that $P_R(0)=\frac{460}{804}\approx 0.572$, $P_C(0)=\frac{133}{804}\approx 0.165$, and $P_T(0)=\frac{211}{804}\approx 0.263$. So $P_R(n)\longrightarrow 0.549$, $P_C(n)\longrightarrow 0.292$, and $P_T(n)\longrightarrow 0.159$. Simulations corresponding to this example are given in Fig. 3.

5 Conclusion

In this document, a formulation of the non symmetric discrete model of size 2×2 has been proposed, which studies a DNA evolution taking into account purines and pyrimidines mutation rates. A simulation

has been performed, to compare the proposal to the well known Jukes and Cantor model. Then all non-symmetrical models of size 3x3 that have 6 parameters have been studied theoretically. They have been tested with numerical simulations, to make a distinction between cytosines and thymines in the former proposal. These two models still remain generic, and can be adapted to a large panel of applications, replacing either the couple (purines, pyrimidines) or the tuple (purines, cytosines, thymines) by any categories of interest.

Remark that the $ura3$ gene is not the unique example of a DNA sequence of interest such that none of the existing nucleotides evolution models cannot be applied due to a complex mutation matrix. For instance, a second gene called $can1$ has been studied too by the authors of [12]. Similarly to gene $ura3$, usual models cannot be used to predict the evolution of $can1$, whereas a study following a same canvas than what has been proposed in this research work can be realized.

In future work, biological consequences of the results produces by these models will be systematically investigated. Then, the most general non symmetric model of size 4 will be regarded in some particular cases taken from biological case studies, and the possibility of mutations non uniformly distributed will then be regarded. Finally, this 4×4 general case will be investigated using Perron-Frobenius based approaches instead of using methods directly inspired by linear algebra, in order to obtain the most global results on mutation matrices.

Competing Interests

The authors declare that no competing interests exist.

References

[1] Jukes TH, Cantor CR. Evolution of protein molecules. Academy Press; 1969.

[2] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 1980;16:111–120. 10.1007/BF01731581.

[3] Felsenstein J. A view of population genetics. Science. 1980;208(4449):1253.

[4] Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. J Mol Evol. 1985;22(2):160-174.

[5] Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and g+c-content biases. Molecular Biology and Evolution. 1992;9(4):678-687.

[6] Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution. 1993;10(3):512-526.

[7] Yang Z. Estimating the pattern of nucleotide substitution. Journal of Molecular Evolution. 1994;10:105-111.

[8] Klosterman, Peter S, Uzilov, Andrew V, Bendaa, Yuri R, Bradley, Robert K, Chao, Sharon and Kosiol, Carolin and Goldman, Nick and Holmes, Ian. XRate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics. 2006;7:428.

[9] Boussau, Gouy. Efficient likelihood computations with nonreversible models of evolution. Systematic Biology. 2006;55(5):756-768.

[10] Yap B, Speed T. Rooting a phylogenetic tree with nonreversible substitution models. BMC Evolutionary Biology. 2005;5(2).

[11] Bahi JM, Guyeux C, Perasso A. Predicting the evolution of gene ura3 in the yeast saccharomyces cerevisiae. Procedia Computer Science. 2012;11:4-16.

[12] Lang GI, Murray AW. Estimating the per-base-pair mutation rate in the yeast saccharomyces cerevisiae. Genetics. 2008;178(1):67-82.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=728&id=6&aid=6913